TEACHING STATISTICS WITH SIMULATED DATA

Andrej Blejec, National Institute of Biology, Slovenia

Real life data are usually used in statistics teaching. For illustration of statistical methods, simulated data with known statistical properties are more suitable. Computer simulations and examples on simulated data can sometimes be used instead of proofs. When the methods are tested and familiarized on data with known properties, they can be effectively used on real life data with unknown Structure and properties.

INTRODUCTION

The aim of this paper is to present some aspects of the use of simulated data in statistics teaching, which enable us to explain and understand statistical concepts.

To present statistics as a useful and interesting topic, teachers usually use interesting real life examples and data. Data with context are used to attract students and make them think about statistical results not only *per se*, but also to derive implications in the sense of given context. There are many benefits of context data as opposed to invented data. Statistical analysis must address some substantive question about data and try to make an answer or statement about that question. Invented data are usually used just to drill the computational aspect of statistics, leaving no room for such questions and answers. Sometimes, in desire to be as realistic as possible, real life examples are too complicated and the meaning of statistical results is not easy to depict. If the example is too complicated, students don’t have enough knowledge and skills to see, which characteristics of data or process are exposed by statistical result.

On the other hand, there are a lot of concepts and methods that should be proved. Most of our students are not mathematically inclined. In such situation, one should think about Moore’s question: *If an audience is not convinced by proof, why do proof?* (Moore, 1997, p.3). Computer simulation, with good data visualisation, is usually more convincing than proofs. They provide a good way, not to prove, but to show what is the meaning of certain statistical method, how it functions and to expose it’s limitations.

LEARNING TO USE TOOLS AND INSTRUMENTS

If we want to use some tool or instrument, we must train ourselves to use it. We can not use the tool, or instrument, on complex objects from the very beginning. For efficient use we need a lot of experiences. At the beginning they are usually gained on some simple, training object. Afterwards, we are able to get new skills and knowledge in more realistic situations. Let us imagine, that we want to find out the structure of a wall in a house. One possible method is to drill a hole into the wall (see Figure 1). As a response, which will show the structure, we will use the colour of material that will be drilled out. At the beginning, something whitish comes out and then something reddish. Inexperienced user can say only that, while experienced one can tell that there is some paint and then there is a brick in the wall. After some drilling, with red outcome, there is suddenly no response. Inexperienced user would say, that we are through, while the experienced one would suspect, that there is a hollow brick in the wall. To support this idea, he would include another response into the "wall analysis", namely the resistance of material to drilling. Since no material comes out, inexperienced user would give up, while the experienced one would analyse the resistance and try to determine hardness of the wall structure which is different for bricks and concrete. As a final response, showing that

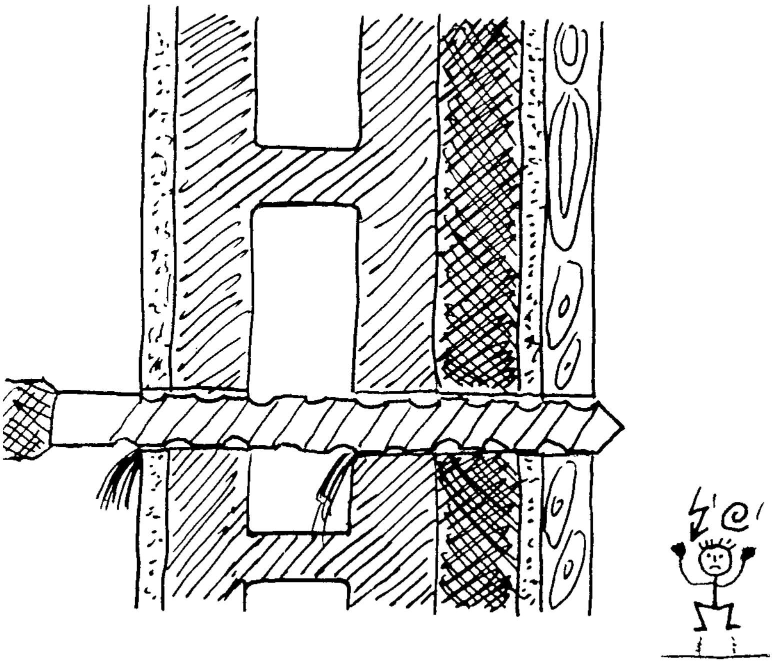

Figure 1. The "wall structure analysis"

we finally drilled through the wall, a third variable can be included, namely the angry neighbour complaining that we drilled a hole into his furniture. How can one get experiences? Certainly not by drilling into the walls with unknown structure. One must take simple objects, with known structure, such as bricks or pieces of concrete, wood and similar and train himself to interpret the response of use of the tool.

A more serious example of such training is from the field of astrophysics. Spectrographs are used to analyse the structure of stellar atmosphere. The spectrum of the Sun is shown in the bottom panel of Figure 2, and one has to tell, what is the composition of the Sun’s atmosphere. A difficult and complex problem for untrained person. The standard procedure of getting experiences for interpretation is, to look at the spectra of known substances and to compare their spectral lines with the lines of more complex one. In the upper panel of Figure 2, one can find the spectrum of the hydrogen. Comparing those two spectra, one can easily recognise hydrogen spectral lines and say that, among other substances, there is hydrogen in the stellar atmosphere. By learning the outcome of spectral analysis in the simple case with known properties, one can recognise the pattern in more complex case.

Statistical methods can be treated as precious instruments which we are trying to learn to use. Too many times, students of statistics are exposed only to complex problems, without training on simple data with known statistical properties - simulated data.

COMPUTER SIMULATED DATA

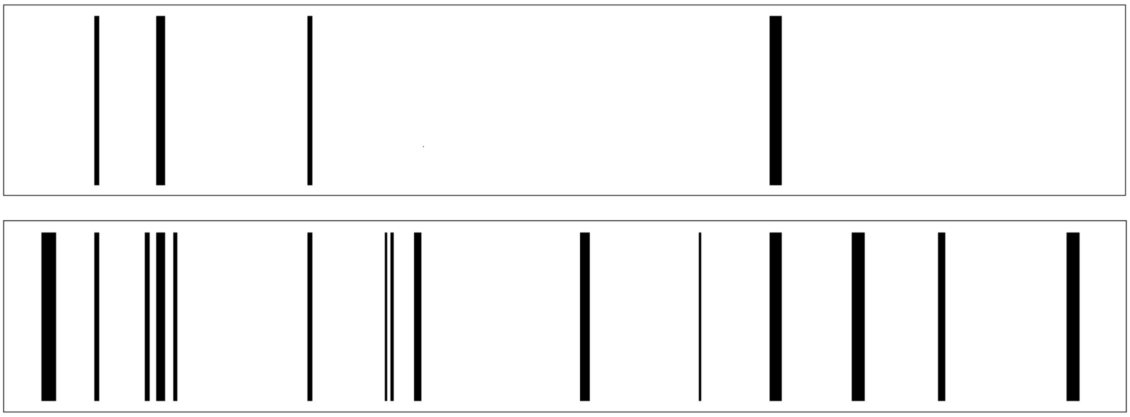

Figure 2. Absorption spectrum of hydrogen (upper panel) and the Sun (lower panel)

Computer simulated data are data, generated as a sample from the population with predefined distribution and parameters. An example can be a sample of certain size, taken from the normal distribution with predefined mean and variance. Such data can be used as a training set, on which methods are examined. Different estimators for parameters give different estimates, which can be compared to *known* value of the parameter that was used to generate data. By repeating the sampling and application of methods to many samples, some feeling can be gained about the quality of estimates. Computer simulations can help us to explain concepts such as confidence intervals, standard error. One can test the behaviour of the method in cases where assumptions, for example random sampling, are not met and show the bias introduced into estimation of the mean by rejecting the maximum of a sample.

One can generate data according to some model and show students the results of statistical analysis. By variation of parameters, and observation of results under different conditions, students will be able to recognise certain characteristics also in complicated, real life problem. For example, data for a linear model $y=\alpha+\beta x+\varepsilon$ can be generated if one defines constants α and β, selects distribution and parameters of x and decides about variability of ε. Various data sets can be generated and graphically presented to students to see the variety of possible patterns generated under same conditions. Different methods for estimation of α and β can be used and their effectiveness to reveal or not to reveal known values of α and β can be shown. In such situation, students can judge, if the method is able to find estimates that are close to the true value or not, what is the influence of the sample size or variability of error term on the effectiveness of estimation. On simulated data they can learn, what they can expect from certain statistical method, what happens if assumptions and restrictions are not met and so on. Later, when analysing real life problems, they will be able not to blindly rely on results but to tell if such results are possible if the conditions and assumptions are met, the model is valid and so on. An example of such data, generated according to the linear model and $r^2 = 0.2$, is shown on Figure 3. In the right panel, a scatterplot with plotted regression line (thin line) and model line (thick line) is plotted. In the right panel, a collection of regression lines, obtained by repeated sampling under the same conditions is plotted. Comparing the set of thin lines with the model line (thick line) one can see that, in majority, we are quite close to the real situation, but in some cases we completely miss the underlying situation. Varying the error term variation (changing r^2) and modifying the sample size, one can get feeling for reliability of linear regression.

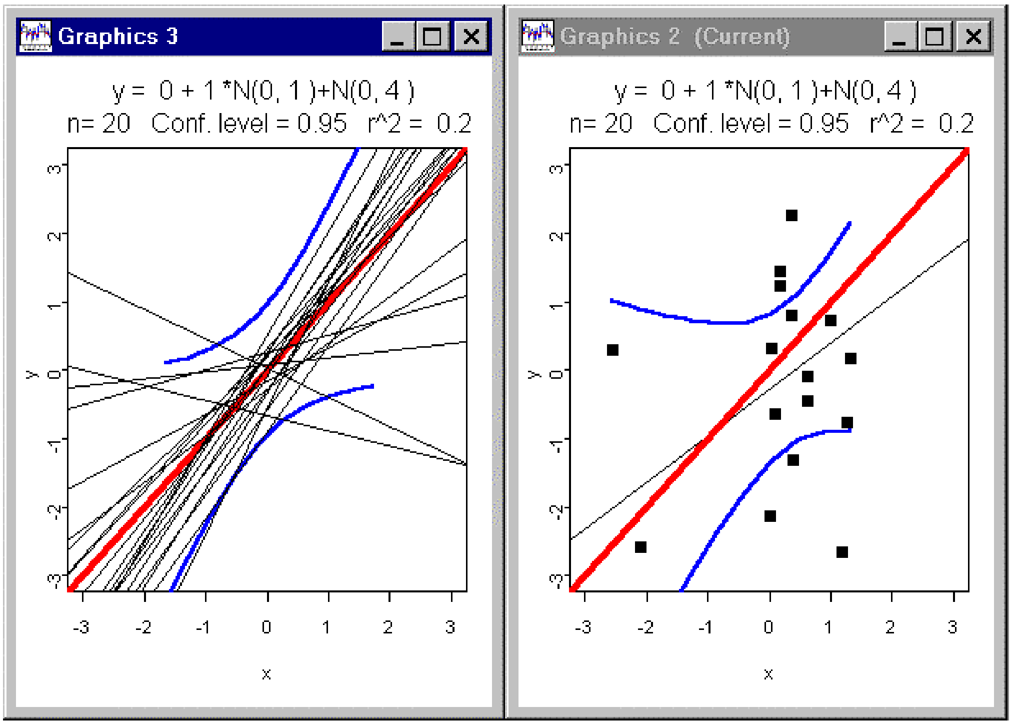
Graphics 3
y = 0 + 1 *N(0, 1)+N(0, 4)
n= 20 Conf. level = 0.95 r^2 = 0.2
y
-3
-2
-1
0
1
2
3
x
Graphics 2 (Current)
y = 0 + 1 *N(0, 1)+N(0, 4)
n= 20 Conf. level = 0.95 r^2 = 0.2
y
-3
-2
-1
0
1
2
3
x

Figure 3. Linear regression: *y=x+ε ; x*~N(0,1) , *ε*~N(0,4)

To illustrate hypothesis testing, one can generate a series of samples, assuming that hypothesis is true. Such series of samples (plotted in left side of panels in Figure 4) shows the diversity of possible samples under the same hypothesis. Students can thus expect what would the sample, taken from an hypothetical population, look like. One can

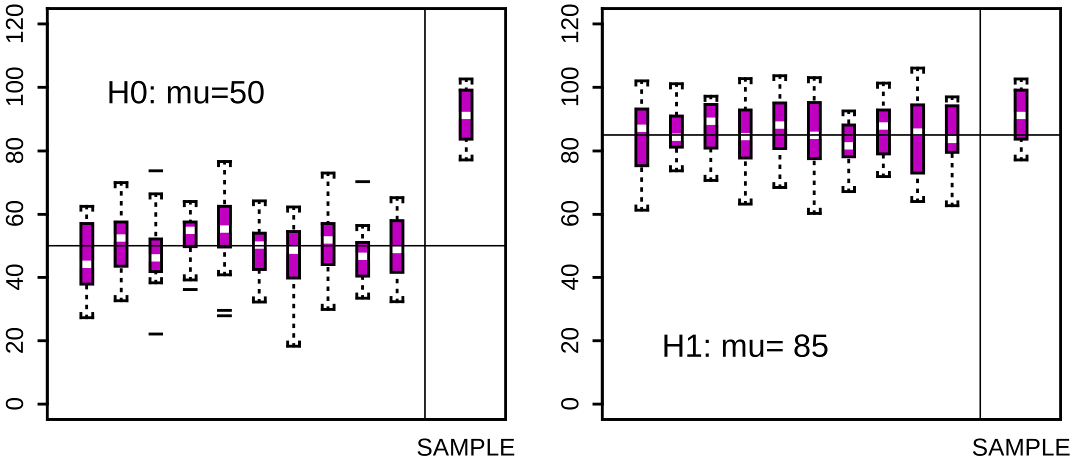

Figure 4. Hypothesis testing. The sample was generated from N(90,100)

compare the sample under consideration with the series of possible ones and try to judge if the sample is similar to possible ones or not. While the situation in left panel of Figure 4 is quite clear, it is not easy to tell, that the considered sample in the right panel was generated from the population with different mean as assumed. For greater impression, one can plot samples generated under hypothesis by replotting them on the same place on the screen.

CONCLUSION

Computer simulated data with known statistical properties, *i.e.* distribution type and parameters, can substantially help us to explain and understand principles of statistical methods. Their greatest value is, that one can compare results of statistical analyses with known situation. One can deliberately violate assumptions and see the effect and susceptibility of results to unfulfilled assumptions. For greater efficiency, specialised teaching and simulation software is needed (Blejec, 1996). Analysing simulated data, students can see results under different conditions and situations. After such training, they can cope with complex real life data, where they must make conclusions about situation from the results. Simulated and real life data are not rivals but companions. Simulated data should be used for illustration of strength of the method and real life data to show practical value and use of statistical reasoning.

REFERENCES

Blejec A. (1996). Using S-PLUS as a Platform for Statistical Simulations and Experiments in Teaching Statistics. In: F Faulbaum and W Bandilla (Eds.) SoftStat'95, Stuttgart: Lucius and Lucius, 373-380.

Moore S. D. (1996). New Pedagogy and New Content: The Case of Statistics. In: Phillips B (Ed.) Papers on Statistical Education. ICME-8, Swinburne University of Technology, 1-4.